Analyzing and modeling European R&D collaborations: Challenges and opportunities from a large social network

Michael J. Barber, Manfred Paier, and Thomas Scherngell

Austrian Research Centers—ARC, Division systems research,
Donau-City-Straße 1, 1220 Vienna, Austria

April 9, 2010

1 Introduction

Networks have attracted a burst of attention in the last decade (useful reviews include references [1, 9, 15, 28]), with applications to natural, social, and technological systems. While networks provide a powerful abstraction for investigating relationships and interactions, the preparation and analysis of complex real-world networks nonetheless presents significant challenges. In particular social networks are characterized by a large number of different properties and generation mechanisms which require a rich set of indicators. The objective of the current study is to analyze large social networks with respect to their community structure and mechanisms of network formation. As a case study, we consider networks derived from the European Union's Framework Programs (FPs) for Research and Technological Development.

The EU FPs were implemented to follow two main strategic objectives: First, strengthening the scientific and technological bases of European industry to foster international competitiveness and, second, the promotion of research activities in support of other EU policies. In spite of their different scopes, the fundamental rationale of the FPs has remained unchanged. All FPs share a few common structural key elements. First, only projects of limited duration that mobilize private and public funds at the national level are funded. Second, the focus of funding is on multinational and multi-actor collaborations that add value by operating at the European level. Third, project proposals are to be submitted by self-organized consortia and the selection for funding is based on specific scientific excellence and socio-economic relevance criteria [33]. By considering the constituents of these consortia, we can represent and analyze the FPs as networks of projects and organizations. The resulting networks are of substantial size, including over 50 thousand projects and over 30 thousand organizations.

We have general interest in studying a real-world network of large size and high complexity from a methodological point of view. Furthermore, socio-economic research emphasizes the central importance of collaborative activities

in R&D for economic competitiveness (see, for instance, reference [16], among many others). Mainly for reasons of data availability, attempts to evaluate quantitatively the structure and function of the large social networks generated in the EU FPs have begun only in the last few years, using social network analysis and complex networks methodologies [2, 6–8, 34]. Studies to date point to the presence of a dense and hierarchical network. A highly connected core of frequent participants, taking leading roles within consortia, is linked to a large number of peripheral actors, forming a giant component that exhibits the characteristics of a small world.

We augment the earlier studies by applying a battery of methods to the most recent data. We begin with constructing the network, discussing needed processing of the raw data in section 2 and continuing with the network definition in section 3. We next examine the overall network structure in section 4, showing that the networks for each FP feature a giant component with highly skewed degree distribution and small world properties. We follow this with an exploration of community structure in sections 5 and 6, showing that the networks are made of heterogeneous subcommunities with strong topical differentiation. Finally, we investigate determinants of network formation with a binary choice model in section 7; this is similar to a recent analysis of Spanish firms [4], but with a focus on the European level and on geographic and network effects. Results are summarized in section 8.

2 Data Preparation

We draw on the latest version of the sysres EUPRO database. This database includes all information publicly available through the CORDIS projects database[1] and is maintained by ARC systems research (ARC sys). The sysres EUPRO database presently comprises data on funded research projects of the EU FPs (complete for FP1–FP5, and about 70% complete for FP6) and all participating organizations. It contains systematic information on project objectives and achievements, project costs, project funding and contract type, as well as information on the participating organizations including the full name, the full address and the type of the organization.

For purposes of network analyses, the main challenge is the inconsistency of the raw data. Apart from incoherent spelling in up to four languages per country, organizations are labelled inhomogeneously. Entries may range from large corporate groupings, such as EADS, Siemens and Philips, or large public research organizations, like CNR, CNRS and CSIC, to individual departments and labs.

Due to these shortcomings, the raw data is of limited use for meaningful network analyses. Further, any fully automated standardization procedure is infeasible. Instead, a labor-intensive, manual data-cleaning process is used in building the database. The data-cleaning process is described in reference [34]; here, we restrict discussion to the steps of the process relevant to the present work. These are:

1. Identification of unique organization name. Organizational boundaries are defined by legal control. Entries are assigned to appropriate organizations

[1] http://cordis.europa.eu

using the more recently available organization name. Most records are easily identified, but, especially for firms, organization names may have changed frequently due to mergers, acquisitions, and divestitures.

2. Creation of subentities. This is the key step for mitigating the bias that arises from the different scales at which participants appear in the data set. Ideally, we use the actual group or organizational unit that participates in each project, but this information is only available for a subset of records, particularly in the case of firms. Instead, subentities that operate in fairly coherent activity areas are pragmatically defined. Wherever possible, subentities are identified at the second lowest hierarchical tier, with each subentity comprising one further hierarchical sub-layer. Thus, universities are broken down into faculties/schools, consisting of departments; research organizations are broken down into institutes, activity areas, etc., consisting of departments, groups or laboratories; and conglomerate firms are broken down into divisions, subsidiaries, etc. Subentities can frequently be identified from the contact information even in the absence of information on the actual participating organizational unit. Note that subentities may still vary considerably in scale.

3. Regionalization. The data set has been regionalized according to the European Nomenclature of Territorial Units for Statistics (NUTS) classification system[2], where possible to the NUTS3 level. Mostly, this has been done via information on postal codes.

Due to resource limitations, only organizations appearing more than thirty times in the standardization table for FP1–FP5 have thus far been processed. This could bias the results; however, the networks have a structure such that the size of the bias is quite low (see reference [34]).

Additionally, we make use of a representative survey[3] of FP5 participants[4]. The survey focuses on the issues of partner selection, intra-project collaboration, and output performance of EU projects on the level of bilateral partnerships, including individuals as well as organizations. As the survey was restricted to small collaborative projects (specifically, projects with a minimum of two and a maximum of 20 partners), the survey addresses a subset of 9,107 relevant (59% of all FP5) projects. It yielded 1,686 valid responses, representing 3% of all (relevant) participants, and covering 1,089 (12% of all relevant) projects.

3 Network Definition

Using the sysres EUPRO database, for each FP we construct a network containing the collaborative projects and all organizational subentities that are

[2]NUTS is a hierarchical system of regions used by the statistical office of the European Community for the production of regional statistics. At the top of the hierarchy are NUTS-0 regions (countries) below which are NUTS-1 regions and then NUTS-2 regions, etc.

[3] This survey was conducted in 2007 by the Austrian Research Centers GmbH, Vienna, Austria and operated by b-wise GmbH, Karlsruhe, Germany.

[4]We chose FP5 (1998-2002) for the survey, in order to cover some of the developments over time, including prior as well as subsequent bilateral collaborations, and effects of the collaboration both with respect to scientific and commercial outcome. Thus, the survey is able to complement the sysres EUPRO database.

participants in those projects. An organization is linked to a project if and only if the organization is a member of the project. Since an edge never exists between two organizations or two projects, the network is bipartite. The network edges are unweighted; in principle, the edges could be assigned weights to reflect the strength of the participation, but the data needed to assign the network weights is not available.

We will also consider, for each FP, the projections of the bipartite networks onto unipartite networks of organizations and projects. The organization projections are constructed by taking the organizations as the vertices, with edges between any organizations that are at distance two in the corresponding bipartite network. Thus, organizations are neighbors in the projection network if they take part in one or more projects together. The project projections are similar, with projects vertices linked when they have one or more participants in common. While the construction of the projection networks intrinsically loses information available in the bipartite networks, they can nonetheless be useful.

For the binary choice model, we construct another network using cross-section data on 191 organizations that are selected from the survey data. We employ the collaboration network of the respondents on the organization level (this network comprises 1,173 organizations collaborating in 1,089 projects) and extract the 2-core [14] of its largest component (203 organizations representing 17% of all vertices)[5]. Finally, another 12 organizations are excluded due to non-availability of geographical distance data, so that we end up with a sample of 191 organizations.

4 Network Structure

We first consider the bipartite networks for each of the FP networks. Call the size of an organization the number of projects in which it takes part, and similarly call the size of a project the number of constituent organizations taking part in the project. These sizes correspond directly to the degrees of the relevant vertices in the bipartite networks. Both parts—organizations (fig. 1) and projects (fig. 2)—of each of the networks feature strongly skewed, heavy tailed size distributions. The sizes of vertices can differ by orders of magnitude, pointing towards the existence of high degree hubs in the networks; hubs of this sort can play an important role in determining the network structure.

The organization size distributions are similar for each of the FPs. The underlying research activities thus have not altered the mix of organizations participating in a particular number of projects in each Framework Program, despite changes in the nature of those research activities over time. In contrast, the rule changes in FP6 that favor larger project consortia are clearly seen in the project size distributions.

Turning to the projection networks, we see that both the organization projection (table 1) and the project projection (table 2) show small-world properties [37]. First, note that the great majority of the N vertices and M edges are in the largest connected component of the networks. In light of this, we focus on paths in only the largest component. The average path length l in each pro-

[5]This technical trick ensures optimal utilization of observed collaborations in the estimation model, while keeping the size of the model small. It is important to note that it does not make use of the network properties on this somewhat arbitrary sub-network.

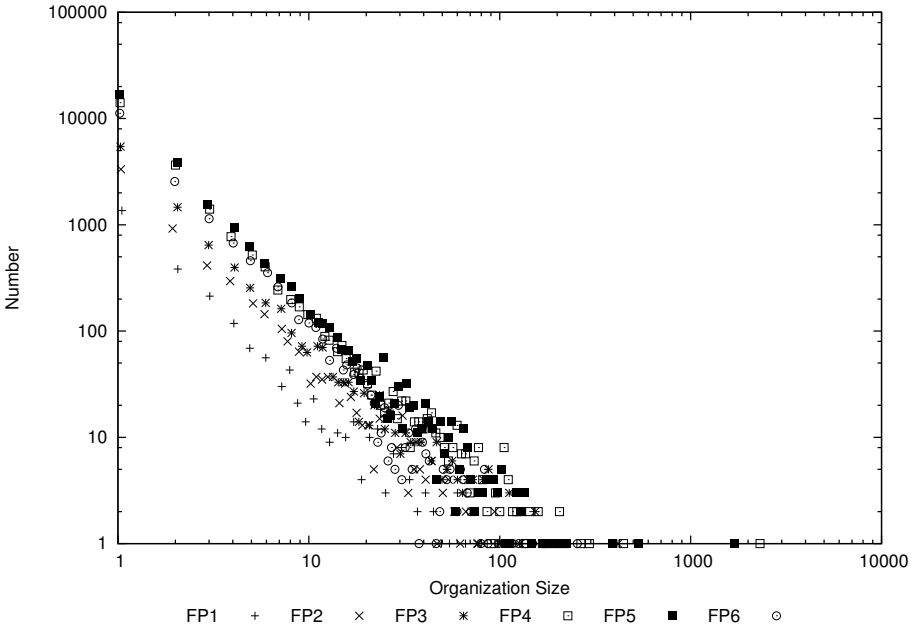

Figure 1: Organization sizes.

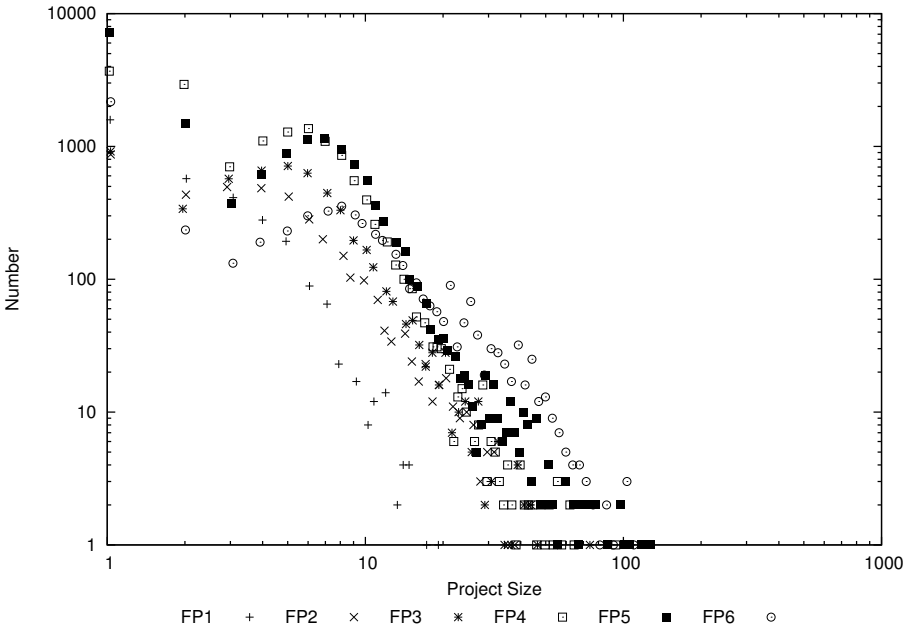

Figure 2: Project sizes.

Table 1: Organization projection properties.

Measure	FP1	FP2	FP3	FP4	FP5	FP6
No. of vertices N	2116	5758	9035	21599	25840	17632
No. of edges M	9489	62194	108868	238585	385740	392879
No. of components	53	45	123	364	630	26
N for largest component	1969	5631	8669	20753	24364	17542
Share of total (%)	93.05	97.79	95.95	96.08	94.29	99.49
M for largest component	9327	62044	108388	237632	384316	392705
Share of total (%)	98.29	99.76	99.56	99.60	99.63	99.96
N for 2nd largest component	8	6	9	10	12	9
M for 2nd largest component	44	30	72	90	132	72
Diameter of largest component	9	7	8	11	10	7
l largest component	3.62	3.21	3.27	3.45	3.30	3.03
Clustering coefficient	0.65	0.74	0.74	0.78	0.76	0.80
Mean degree	9.0	21.6	24.1	22.1	29.9	44.6
Fraction of N above the mean (%)	29.4	28.0	23.6	22.4	23.5	26.1

jection network is short, as is the diameter. However, the clustering coefficient [37], which ranges between zero and one, is high. The combination of short path length and high clustering is characteristic of small world networks. The small-world character is expected to be beneficial in the FP networks, as small-world networks have been shown to encourage the spread of knowledge in model systems [12].

Additionally, the heavy tailed size distributions of the bipartite networks has a visible effect on the degrees of the projection networks. In each case, the data is quite asymmetric about the mean degree, as seen by examining what fraction of vertices have degree above the mean. The fractions are between 20% and 30%, consistent with the skewed degree distributions (the distributions are shown in references [6, 34]; the relation between the degrees in the bipartite networks and the projections is explored in reference [6]).

5 Community Structure

Of great current interest is the identification of community groups, or modules, within networks. Stated informally, a community group is a portion of the network whose members are more tightly linked to one another than to other members of the network. A variety of approaches [3, 11, 19, 20, 22, 26, 27, 30, 32] have been taken to explore this concept; see references [13, 25] for useful reviews. Detecting community groups allows quantitative investigation of relevant subnetworks. Properties of the subnetworks may differ from the aggregate properties of the network as a whole, e.g., modules in the World Wide Web are sets of topically related web pages. Thus, identification of community groups within a network is a first step towards understanding the heterogeneous substructures of the network.

Methods for identifying community groups can be specialized to distinct classes of networks, such as bipartite networks [5, 21]. This is immediately relevant for our study of the FP networks, allowing us to examine the community

Table 2: Project projection properties.

Measure	FP1	FP2	FP3	FP4	FP5	FP6
No. of vertices N	2116	5758	9035	21599	25840	17632
No. of edges M	9489	62194	108868	238585	385740	392879
No. of components	53	45	123	364	630	26
N for largest component	1969	5631	8669	20753	24364	17542
Share of total (%)	93.05	97.79	95.95	96.08	94.29	99.49
M for largest component	9327	62044	108388	237632	384316	392705
Share of total (%)	98.29	99.76	99.56	99.60	99.63	99.96
N for 2nd largest component	8	6	9	10	12	9
M for 2nd largest component	44	30	72	90	132	72
Diameter of largest component	9	7	8	11	10	7
l largest component	3.62	3.21	3.27	3.45	3.30	3.03
Clustering coefficient	0.65	0.74	0.74	0.78	0.76	0.80
Mean degree	9.0	21.6	24.1	22.1	29.9	44.6
Fraction of N above the mean (%)	29.4	28.0	23.6	22.4	23.5	26.1

structure in the bipartite networks. Communities are expected to be formed of groups of organizations engaged in R&D into similar topics, and the projects in which those organizations take part.

5.1 Modularity

To identify communities, we take as our starting point the modularity, introduced by Newman and Girvan [26]. Modularity makes intuitive notions of community groups precise by comparing network edges to those of a null model. The modularity Q is proportional to the difference between the number of edges within communities c and those for a null model:

$$Q \equiv \frac{1}{2M} \sum_{c} \sum_{i,j \in c} (A_{ij} - P_{ij}) \quad . \tag{1}$$

Along with eq. (1), it is necessary to provide a null model, defining P_{ij}.

The standard choice for the null model constrains the degree distribution for the vertices to match the degree distribution in the actual network. Random graph models of this sort are obtained [10] by putting an edge between vertices i and j at random, with the constraint that on average the degree of any vertex i is d_i. This constrains the expected adjacency matrix such that

$$d_i = E\left(\sum_j A_{ij}\right) \quad . \tag{2}$$

Denote $E(A_{ij})$ by P_{ij} and assume further that P_{ij} factorizes into

$$P_{ij} = p_i p_j \quad , \tag{3}$$

leading to

$$P_{ij} \equiv \frac{d_i d_j}{2M} \quad . \tag{4}$$

A consequence of the null model choice is that $Q = 0$ when all vertices are in the same community.

The goal now is to find a division of the vertices into communities such that the modularity Q is maximal. An exhaustive search for a decomposition is out of the question: even for moderately large graphs there are far too many ways to decompose them into communities. Fast approximate algorithms do exist (see, for example, references [24, 31]).

5.2 Finding Communities in Bipartite Networks

Specific classes of networks have additional constraints that can be reflected in the null model. For bipartite graphs, the null model should be modified to reproduce the characteristic form of bipartite adjacency matrices:

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{O} & \boldsymbol{M} \\ \boldsymbol{M}^{\mathrm{T}} & \boldsymbol{O} \end{bmatrix} \quad . \tag{5}$$

Recently, specialized modularity measures and search algorithms have been proposed for finding communities in bipartite networks [5, 21]. These measures and methods have not been studied as extensively as the versions with the standard null model shown above, but many of the algorithms can be adapted to the bipartite versions without difficulty. Limitations of modularity-based methods (e.g., the resolution limit described in reference [17]) are expected to hold as well.

We make use of the algorithm called BRIM: bipartite, recursively induced modules [5]. BRIM is a conceptually simple, greedy search algorithm that capitalizes on the separation between the two parts of a bipartite network. Starting from some partition of the vertices of type 1, it is straightforward to identify the optimal partition of the vertices of type 2. From there, optimize the partition of vertices of type 1, and so on. In this fashion, modularity increases until a (local) maximum is reached. However, the question remains: is the maximum a "good" one? At this level then a random search is called for, varying the composition and number of communities, with the goal of reaching a better maximum after a new sequence of searching using the BRIM algorithm.

6 Communities in the Framework Program Networks

A popular approach in social network analysis—where networks are often small, consisting of a few dozen nodes—is to visualize the networks and identify community groups by eye. However, the Framework Program networks are much larger: can we "see" the community groups in these networks?

Structural differences or similarities of such networks are not obvious at a glance. For a graphical representation of the organizations and/or projects by dots on an A4 sheet of paper, we would need to put these dots at a distance of about 1 mm from each other, and we then still would not have drawn the links (collaborations) which connect them.

Previous studies used a list of coarse graining recipes to compact the networks into a form which would lend itself to a graphical representation [8].

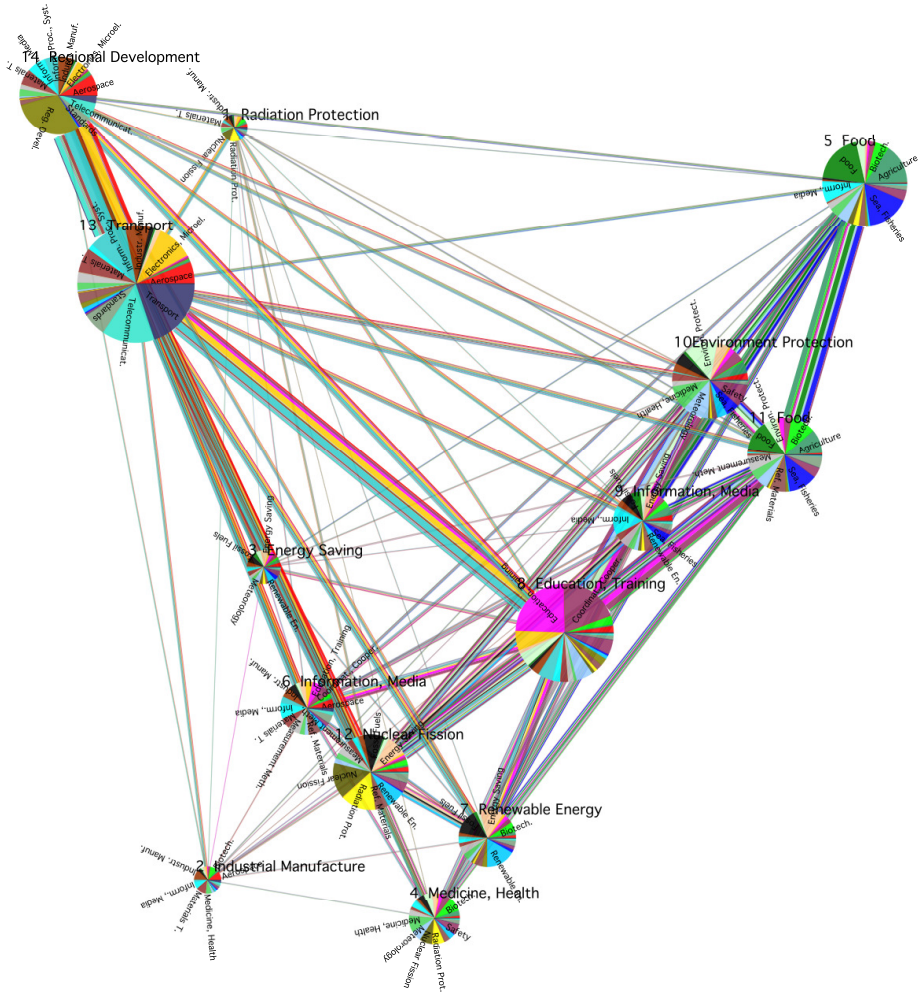

Figure 3: Community groups in the network of projects and organizations for FP3.

As an alternative we have attempted to detect communities just using BRIM, i.e., purely on the basis of relational network structure, ignoring any additional information about the nature of agents.

In fig. 3, we show a community structure for FP3 found using the BRIM algorithm, with a modularity of $Q = 0.602$ for 14 community groups. The communities are shown as vertices in a network, with the vertex positions determined using spectral methods [35]. The area of each vertex is proportional to the number of edges from the original network within the corresponding community. The width of each edge in the community network is proportional to the number of edges in the original network connecting community members from the two linked groups. The vertices and edges are shaded to provide additional information about their topical structure, as described in the next section. Each community is labeled with the most frequently occurring subject index.

6.1 Topical Profiles of Communities

Projects are assigned one or more standardized subject indices. There are 49 subject indices in total, ranging from *Aerospace* to *Waste Management*. We denote by

$$f(t) > 0 \tag{6}$$

the frequency of occurrence of the subject index t in the network, with

$$\sum_t f(t) = 1 \quad . \tag{7}$$

Similarly we consider the projects within one community c and the frequency

$$f_c(t) \geq 0 \tag{8}$$

of any subject index t appearing in the projects only of that community. We call f_c the topical profile of community c to be compared with that of the network as a whole.

Topical differentiation of communities can be measured by comparing their profiles, among each other or with respect to the overall network. This can be done in a variety of ways [18], such as by the Kullback "distance"

$$D_c = \sum_t f_c(t) \ln \frac{f_c(t)}{f(t)} \quad . \tag{9}$$

A true metric is given by

$$d_c = \sum_t |f_c(t) - f(t)| \quad , \tag{10}$$

ranging from zero to two.

Topical differentiation is illustrated in fig. 4. In the figure, example profiles are shown, taken from the network in fig. 3. The community-specific profile corresponds to the community labeled '11. Food" in fig. 3. Based on the most frequently occurring subject indices—*Agriculture*, *Food*, and *Resources of the Seas, Fisheries*—the community consists of projects and organizations focussed on R&D related to food products. The topical differentiation is $d_c = 0.90$ for the community shown.

7 Binary Choice Model

We now turn to modeling organizational collaboration choices in order to examine how specific individual characteristics, spatial effects, and network effects determine the choice of collaboration (the theoretical underpinnings are described in reference [29]. We will build upon the survey data and the subnetwork constructed therefrom (section 3). While this restricts us to only 191 organizations, we have considerably more information about these organizations than for the complete networks.

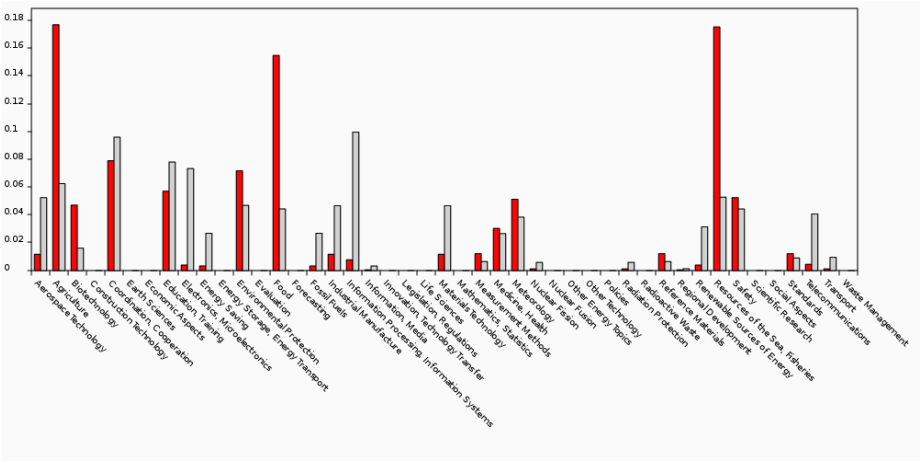

Figure 4: Topical differentiation in a network community. The histogram shows the difference between the topical profile $f_c(t)$ for a specific community (dark bars) and the overall profile $f(t)$ for the network as a whole (light bars). The community-specific profile shown is for the community labeled "11. Food" in fig. 3. The community has $d_c = 0.90$.

7.1 The Empirical Model

In our analytical framework, the constitution of a collaboration Y_{ij} between two organizations i and j will depend on an unobserved continuous variable Y_{ij}^* that corresponds to the profit that two organizations i and j receive when they collaborate. Since we cannot observe Y_{ij}^* but only its dichotomous realizations Y_{ij}, we assume $Y_{ij} = 1$ if $Y_{ij}^* > 0$ and $Y_{ij} = 0$ if $Y_{ij}^* \leq 0$. Y_{ij} is assumed to follow a Bernoulli distribution so that Y_{ij} can take the values one and zero with probabilities π_{ij} and $1 - \pi_{ij}$, respectively. The probability function can be written as

$$\Pr(Y_{ij}) = \pi_{ij}^{Y_{ij}} (1 - \pi_{ij})^{1 - Y_{ij}} \tag{11}$$

with $E[Y_{ij}] = \mu_{ij} = \pi_{ij}$ and $Var[Y_{ij}] = \sigma_{ij}^2 = \pi_{ij}(1 - \pi_{ij})$, where μ_{ij} denotes some mean value.

The next step in defining the model concerns the systematic structure—we would like the probabilities π_{ij} to depend on a matrix of observed covariates. Thus, we let the probabilities π_{ij} be a linear function of the covariates:

$$\pi_{ij} = \sum_{k=1}^{K} \beta_k X_{ij}^{(k)} \quad , \tag{12}$$

where the $X_{ij}^{(k)}$ are elements of the $\boldsymbol{X}^{(k)}$ matrix containing a constant and $K-1$ explanatory variables, including geographical effects, relational effects and FP experience characteristics of organizations i and j. $\boldsymbol{\beta}_K = (\beta_0, \boldsymbol{\beta}_{K-1})$ is the $K \times 1$ parameter vector, where β_0 is a scalar constant term and $\boldsymbol{\beta}_{K-1}$ is the vector of parameters associated with the $K-1$ explanatory variables.

However, estimating this model using ordinary least squares procedures is not convenient since the probability π_{ij} must be between zero and one, while

the linear predictor can take any real value. Thus, there is no guarantee that the predicted values will be in the correct range without imposing any complex restrictions[23]. A very promising solution to this problem is to use the logit transform of π_{ij} in the model, i.e., replacing eq. (12) by the following *ansatz*:

$$Logit\left(\pi_{ij}\right) = \log \frac{\pi_{ij}}{1-\pi_{ij}} = h_{ij} \quad , \tag{13}$$

where we have introduced the abbreviation h_{ij}, defined as

$$h_{ij} = \beta_0 + \beta_1 X_{ij}^{(1)} + \beta_2 X_{ij}^{(2)} + \cdots + \beta_K \quad . \tag{14}$$

This leads to the binary logistic regression model to be estimated given by

$$\Pr\left(Y_{ij} = 1 \mid X_{ij}^{(k)}\right) = \pi_{ij} = \frac{\exp\left(h_{ij}\right)}{1+\exp\left(h_{ij}\right)} \quad . \tag{15}$$

The focus of interest is on estimating the parameters $\boldsymbol{\beta}$. The standard estimator for the logistic model is the maximum likelihood estimator. The reduced log-likelihood function is given by [23]

$$\log L\left(\boldsymbol{\beta} \mid Y_{ij}\right) = -\sum_{i,j} \log\left(1+\exp\left(\left(1-2Y_{ij}\right)h_{ij}\right)\right) \quad , \tag{16}$$

assuming independence over the observations Y_{ij}. The resulting variance matrix $\boldsymbol{V}\left(\hat{\boldsymbol{\beta}}\right)$ of the parameters is used to calculate standard errors. $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically efficient when the observations of Y_{ij} are stochastic and in absence of multicollinearity among the covariates.

7.2 Variable Construction

7.2.1 The Dependent Variable

To construct the dependent variable Y_{ij} that corresponds to observed collaborations between two organizations i and j, we construct the $n \times n$ collaboration matrix $\boldsymbol{Y}$ that contains the collaborative links between the (i, j)-organizations. One element Y_{ij} denotes the existence of collaboration between two organizations i and j as measured in terms of the existence of a common project. $\boldsymbol{Y}$ is symmetric by construction so that $Y_{ij} = Y_{ji}$. Note that the matrix is very sparse. The number of observed collaborations is 702 so that proportion of zeros is about 98%. The mean collaboration intensity between all (i, j)-organizations is 0.02.

7.2.2 Variables Accounting for Geographical Effects

We use two variables $x_{ij}^{(1)}$ and $x_{ij}^{(2)}$ to account for geographical effects on the collaboration choice. The first step is to assign specific NUTS-2 regions to each of the 191 organizations that are given in the sysres EUPRO database. Then we take the great circle distance between the economic centers of the regions where the organizations i and j are located to measure the geographical distance variable $x_{ij}^{(1)}$. The second variable, $x_{ij}^{(2)}$, controls for country border effects and is measured in terms of a dummy variable that takes a value of zero if two organizations i and j are located in the same country, and zero otherwise, in order to get empirical insight on the role of country borders for collaboration choice of organizations.

7.2.3 Variables Accounting for FP Experience of Organizations

This set of variables controls for the experience of the organizations with respect to participation in the European FPs. First, thematic specialization within FP5 is expected to influence the potential to collaborate. We define a measure of thematic distance $x_{ij}^{(3)}$ between any two organizations that is constructed in the following way: Each organization is associated with a unit vector of specialization $\boldsymbol{s}_i$ that relates to the number of project participations $N_{i,1}, \ldots, N_{i,7}$ of organization i in the seven sub-programs of FP5[6].

$$\boldsymbol{s}_i = (N_{i,1}, \ldots, N_{i,7}) \,/ \sqrt{N_{i,1}^2 + \cdots + N_{i,7}^2} \quad . \tag{17}$$

The thematic distance of organizations i and j is then defined as the Euclidean distance of their respective specialization vectors $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, giving $x_{ij}^{(3)} = x_{ji}^{(3)}$ and $0 \leq x_{ij}^{(3)} \leq \sqrt{2}$. The second variable accounting for FP experience focuses on the individual (or research group) level, and takes into account the respondents inclination or openness to FP research. As a proxy for openness of an organization i to FP research, we choose the total number P_i of FP5 projects in the respondent's own organization, that they are aware of[7]. Then we define

$$x_{ij}^{(4)} = P_i + P_i \tag{18}$$

as a measure for the aggregated openness of the respective pair of organizations to FP research. The third variable related with FP experience is the overall number of FP5 project participations an organization is engaged in. Denoting, as above, $N_i = N_{i,1} + \cdots + N_{i,7}$ as the total number of project participations of organization i in FP5, we define

$$x_{ij}^{(5)} = |N_i - N_j| \tag{19}$$

as the difference in the number of participations of organization i and j in FP5. It is taken from the sysres EUPRO database and is an integer ranging from $0 \leq x_{ij}^{(5)} \leq 1{,}228$, resulting from the minimal value of one participation and the maximum of 1,229 participations among the sample of 191 organizations.

7.2.4 Variables Accounting for Relational Effects

We consider a set of three variables accounting for potential relational effects on the decision to collaborate. Hereby, we distinguish between joint history and network effects. The first factor to be taken into account is prior acquaintance of two organizations, and is measured by a binary variable denoting acquaintance on the individual (research group) level before the FP5 collaboration started. It is taken from the survey[8]. By convention, $x_{ij}^{(6)} = 1$ if at least one respondent from organization i nominated organization j as prior acquainted, $x_{ij}^{(5)} = 0$

[6] EESD, GROWTH, HUMAN POTENTIAL, INCO 2, INNOVATION-SME, IST, and LIFE QUALITY

[7] The exact wording of the question was, 'How many FP5 projects of your organization are you aware of?' For multiple responses from an organization, the numbers of known projects are summarized. In cases of missing data, this number is set to zero.

[8] The exact wording of the question was, 'Which of your [project acronym] partners (i.e., persons from which organization) did you know before the project began?'

otherwise. All other relational factors we take into account in the model are network effects. For conceptual reasons we must look at the global FP5 network, where we make use of the structural embeddedness of our 191 sample organizations.

One of the most important centrality measures is betweenness centrality. Betweenness is a centrality concept based on the question to what extent a vertex in a network is able to control the information flow through the whole network [36]. Organizations that are high in betweenness, may thus be especially attractive as collaboration partners. More formally, the betweenness centrality of a vertex can be defined as the probability that a shortest path between a pair of vertices of the network passes through this vertex. Thus, if $B(k,l;i)$ is the number of shortest paths between vertices k and l passing through vertex i, and $B(k,l)$ is the total number of shortest paths between vertices k and l, then

$$b(i) = \sum_{k \neq l} \frac{B(k,l;i)}{B(k,l)} \tag{20}$$

is called the betweenness centrality of vertex i [15]. We calculate the betweenness centralities in the global FP5 network and include

$$x_{ij}^{(7)} = b(i)\, b(j) \tag{21}$$

as a combined betweenness measure.

The third variable accounting for relational effects is local clustering. Due to social closure, we may assume that within densely connected clusters organizations are mutually quite similar, so that it might be strategically advantageous to search for complementary partners from outside. Hereby, communities with lower clustering may be easier to access. We use the clustering coefficient $CC_1(i)$, which is the share of existing links in the number of all possible links in the direct neighborhood (at distance $d = 1$) of a vertex i. Thus, let k_i be the number of direct neighbors and T_i the number of existing links among these direct neighbors, then

$$CC_1(i) = \frac{2T_i}{k_i(k_i - 1)} \tag{22}$$

is the relevant clustering coefficient [37]. We employ the difference in the local clustering coefficients within the global FP5 network for inclusion in the statistical model, by setting

$$x_{ij}^{(8)} = |CC_1(i) - CC_1(j)| \tag{23}$$

in order to obtain a symmetric variable in i and j.

7.3 Estimation Results

This section discusses the estimation results of the binary choice model of R&D collaborations as given by eq. (15). The binary dependent variable corresponds to observed collaborations between two organizations i and j, taking a value of one if they collaborate and zero otherwise. The independent variables are geographical separation variables, variables capturing FP experience of the organizations and relational effects (joint history and network effects). We estimate

Table 3: Maximum likelihood estimation results for the collaboration model based on n^2=36,481 observations. Asymptotic standard errors are given parenthetically.

Coefficient	Basic Model	Extended Model	Full Model
β_0	-1.882***(0.313)	-1.951***(0.342)	-1.816***(0.385)
β_1	-0.145***(0.038)	-0.116***(0.039)	-0.128***(0.040)
β_2	—	-0.103***(0.034)	-0.094**(0.034)
β_3	-1.477***(0.110)	-1.465***(0.114)	-1.589***(0.117)
β_4	—	0.004***(0.001)	0.003***(0.001)
β_5	—	—	0.001(0.000)
β_6	4.224***(0.089)	4.189***(0.089)	4.194***(0.089)
β_7	0.161***(0.023)	0.135***(0.025)	0.119***(0.027
β_8	—	—	0.070**(0.025)

three model versions: The standard model includes one variable for geographical effects and FP experience, respectively, and two variables accounting for relational effects. In the extended model version we add country border effects as additional geographical variable in order to isolate country border effects from geographical distance effects, and openness to FP research as additional FP experience variable. The full model additionally includes balance variables accounting for FP experience and network effects, respectively.

Table 3 presents the sample estimates derived from maximum likelihood estimation for the model versions. The number of observations is equal to 36,481, asymptotic standard errors are given in parentheses. The statistics given in table 4 indicate that the selected covariates show a quite high predictive ability. The Goodman-Kruskal-Gamma statistic ranges from 0.769 for the basic and 0.782 for the extended model to 0.786 for the full model, indicating that more than 75% fewer errors are made in predicting interorganizational collaboration choices by using the estimated probabilities than by the probability distribution of the dependent variable alone. The Somers D statistic and the C index confirm these findings. The Nagelkerke's R-Squared is 0.391 for the basic model, 0.395 for the extended model and 0.397 for the full model version, respectively[9]. A likelihood ratio test for the null hypothesis of $\beta_k = 0$ yields a χ^2_4 test statistic of 2,565.165 for the basic model, a χ^2_6 test statistic of 2,582.421 for the extended model and a χ^2_8 test statistic of 2,597.911 for the full model. These are statistically significant and we reject the null hypothesis that the model parameters are zero for all model versions.

The model reveals some promising empirical insight in the context of the relevant literature on innovation as well as on social networks. The results provide a fairly remarkable confirmation of the role of geographical effects, FP experience effects and network effects for interorganizational collaboration choice in EU FP R&D networks. In general, the parameter estimates are statistically significant and quite robust over different model versions.

[9]Nagelkerke's R-squared is an attempt to imitate the interpretation of multiple R-Squared measures from linear regressions based on the log likelihood of the final model versus log likelihood of the null model. It is defined as $R^2_{\text{Nag}} = \left[1 = (L_0/L_1)^{2/n}\right] / \left[1 - L_0^{2/n}\right]$ where L_0 is the log likelihood of the null model, L_1 is the log likelihood of the model to be evaluated and n is the number of observations.

Table 4: Performance of the three collaboration model versions based on n^2=36,481 observations.

Performance	Basic Model	Extended Model	Full Model
Somers D	0.733	0.746	0.753
Goodman-Kruskal Gamma	0.769	0.782	0.786
C index	0.876	0.873	0.875
Nagelkerke R-squared index	0.391	0.395	0.397
Log-likelihood	-2,190.151	-2,176.768	-2,169.578
Likelihood Ratio Test	2,565.156***	2,582.421***	2,597.911***

The results of the basic model show that geographical distance between two organizations significantly determines the probability to collaborate. The parameter estimate of $\beta_1 = -0.145$ indicates that for any additional 100 km between two organizations the mean collaboration frequency decreases by about 15.6%. Geographical effects matter but effects of FP experience of organizations are more important. As evidenced by the estimate $\beta_3 = -1.477$ it is most likely that organizations choose partners that are located closely in thematic space. A one percent increase in thematic distance reduces the probability of collaboration by more than 3.25%. Most important determinants of collaboration choice are network effects. The estimate of $\beta_6 = 4.224$ tells us that the probability of collaboration between two organizations increases by 68.89% when they are prior acquaintances. Also network embeddedness matters as given by the estimate for $\beta_7 = 0.161$ indicating that choice of collaboration is more likely between organizations that are central players in the network with respect to betweenness centrality.

Turning to the results of the extended model version it can be seen that taking into account country border effects decreases geographical distance effects by about 24% ($\beta_1 = -0.116$). The existence of a country border between two organizations has a significant negative effect on their collaboration probability, the effect is slightly smaller than geographical distance effects ($\beta_2 = -0.103$). Adding openness to FPs as an additional variable to capture FP experience does not influence the other model parameters much. Openness to FPs, though statistically significant, shows only a small impact on collaboration choice.

In the full model version we add one balance variable accounting for FP experience and network effects, respectively. The difference in the number of submitted FP projects has virtually no effect on the choice of collaboration as given by the estimate of β_5. An interesting result from a social network analysis perspective provides the integration of the difference between two organizations with respect to the clustering coefficient. The estimate of $\beta_8 = 0.070$ tells us that it is more likely that two organizations collaborate when the difference of their cluster coefficients is higher. This result points to the existence of strategic collaboration choices for organizations that are highly cross-linked searching for organizations to collaborate with lower clustering coefficients, and the other way round. The effect is statistically significant but smaller than other network effects and geographical effects.

8 Summary

We have presented an investigation of networks derived from the European Union's Framework Programs for Research and Technological Development. The networks are of substantial size, complexity, and economic importance. We have attempted to provide a coherent picture of the complete process, beginning with data preparation and network definition, then continuing with analysis of the network structure and modeling of network formation.

We first considered the challenges involved in dealing with a large amount of imperfect data, detailing the tradeoffs made to clean the raw data into a usable form under finite resource constraints. The processed data was then used to define bipartite networks with vertices consisting of all the projects and organizations involved in each FP. To provide alternative views of the data, we defined projection networks for each part (organizations or projects) of the bipartite networks. Additionally, we used results of a survey of FP5 participants to define a smaller network about which we have more detailed information than we have for the networks as a whole.

Next we examined structural properties of the bipartite and projection networks. We found that the vertex degrees in the FP networks have a highly skewed, heavy tailed distribution. The networks further show characteristic features of small-world networks, having both high clustering coefficients and short average path lengths. We followed this with analysis of the community structure of the Framework Programs. Using a modularity measure and search algorithm adapted to bipartite networks, we identified communities from the networks, and found that the communities are topically differentiated based on the standardized subject indices for Framework Program projects.

In the final stage of analysis, we constructed a binary choice model to explore determinants of inter-organizational collaboration choice. The model parameters were estimated using logistic regression. The model results show that geographical effects matter, but are not the most important determinants. The strongest effect comes from relational characteristics, in particular prior acquaintance, and to a minor extent, network centrality. Also, thematic similarity between organizations highly favors a partnership.

By using a variety of networks and analyses, we have been able to address several different questions about the Framework Programs. The results complement one another, giving a more complete picture of the Framework Programs than the results from any one method alone. We are confident that our understanding of collaborative R&D in the European Union can be improved by extending the analyses presented in this chapter and by expanding the types of analyses we undertake.

Acknowledgments

The authors gratefully acknowledge financial support from the European FP6-NEST-Adventure Program, under contract number 028875 (project NEMO: Network Models, Governance, and R&D Collaboration Networks).

References

[1] Reka Albert and Albert-Laszlo Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002. URL http://link.aps.org/abstract/RMP/v74/p47.

[2] Juan A. Almendral, J. G. Oliveira, L. López, Miguel A. F. Sanjuán, and J. F. F. Mendes. The interplay of universities and industry through the FP5 network. *New Journal of Physics*, 9(6):183–98, 2007. doi: 10.1088/1367-2630/9/6/183. URL http://www.iop.org/EJ/abstract/1367-2630/9/6/183/.

[3] Leonardo Angelini, Stefano Boccaletti, Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. Identification of network modules by optimization of ratio association. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(2):023114, 2007. doi: 10.1063/1.2732162. URL http://arxiv.org/abs/cond-mat/0610182.

[4] N. Arranz and J. C. Fernández de Arroyabe. The choice of partners in R&D cooperation: An empirical analysis of Spanish firms. *Technovation*, 28:88–100, 2008.

[5] Michael J. Barber. Modularity and community detection in bipartite networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 76(6):066102, 2007. doi: 10.1103/PhysRevE.76.066102.

[6] Michael J. Barber, Andreas Krueger, Tyll Krueger, and Thomas Roediger-Schluga. Network of European Union-funded collaborative research and development projects. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 73(3):036132, 2006. doi: 10.1103/PhysRevE.73.036132. URL http://link.aps.org/abstract/PRE/v73/e036132.

[7] Michael J. Barber, Margarida Faria, Ludwig Streit, and Oleg Strogan. Searching for communities in bipartite networks. In Christopher C. Bernido and M. Victoria Carpio-Bernido, editors, *Proceedings of the 5th Jagna International Workshop: Stochastic and Quantum Dynamics of Biomolecular Systems*, New York, NY, 2008. Springer.

[8] S. Breschi and L. Cusmano. Unveiling the texture of a European Research Area: Emergence of oligarchic networks under EU Framework Programmes. *International Journal of Technology Management*, 27(8):747–72, 2004.

[9] Claire Christensen and Reka Albert. Using graph concepts to understand the organization of complex systems. *International Journal of Bifurcation and Chaos*, 17(7):2201–2214, 2007. URL http://arxiv.org/abs/q-bio.OT/0609036. Special Issue "Complex Networks' Structure and Dynamics".

[10] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145, 2002.

[11] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 70(6):066111, 2004. URL http://link.aps.org/abstract/PRE/v70/e066111.

[12] Robin Cowan and Nicolas Jonard. Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control*, 28(8):1557–1575, June 2004. doi: 10.1016/j.jedc.2003.04.002. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V85-4B3K2R3-2&_coverDate=06%2F30%2F2004&_alid=490156249&_rdoc=1&_fmt=&_orig=search&_qd=1&_cdi=5861&_sort=d&view=c&_acct=C000050324&_version=1&_urlVersion=0&_userid=1013681&md5=03425d831627fe203e81452bc8da30dc.

[13] Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *J. Stat. Mech.*, page P09008, 2005. doi: 10.1088/1742-5468/2005/09/P09008. URL http://www.iop.org/EJ/article/1742-5468/2005/09/P09008/jstat5_09_p09008.html.

[14] W. deNooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2004.

[15] S. N. Dorogovtsev and J. F. F. Mendes. The shortest path to complex networks. In N. Johnson, J. Efstathiou, and F. Reed-Tsochas, editors, *Complex Systems and Inter-disciplinary Science*. World Scientific, 2004. URL http://arxiv.org/abs/cond-mat/0404593.

[16] J. Fagerberg, D.C. Mowery, and R.R. Nelson. *The Oxford Handbook of Innovation*. Oxford University Press, 2005.

[17] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *PNAS*, 104(1):36–41, 2007. doi: 10.1073/pnas.0605965104. URL http://www.pnas.org/cgi/reprint/104/1/36.pdf.

[18] A.L. Gibbs and F.E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[19] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002. URL http://www.pnas.org/cgi/content/abstract/99/12/7821.

[20] V. Gol'dshtein and G. A. Koganov. An indicator for community structure. Preprint, July 2006. URL http://arxiv.org/abs/physics/0607159.

[21] Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral. Module identification in bipartite and directed networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 76(3):036102, 2007. doi: 10.1103/PhysRevE.76.036102. URL http://link.aps.org/abstract/PRE/v76/e036102.

[22] Matthew B. Hastings. Community detection as an inference problem. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74 (3):035102, 2006. doi: 10.1103/PhysRevE.74.035102. URL http://arxiv.org/abs/cond-mat/0604429.

[23] J. Johnston and J. Dinardo. *Econometric Methods*. McGraw-Hill, Inc., New York, NY, USA, 2007.

[24] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 69(6):066133, 2004. URL http://link.aps.org/abstract/PRE/v69/e066133.

[25] M. E. J. Newman. Detecting community structure in networks. *Eur. Phys. J. B*, 38:321–330, 2004. URL http://www-personal.umich.edu/~mejn/papers/epjb.pdf.

[26] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 69(2):026113, 2004. URL http://link.aps.org/abstract/PRE/v69/e026113.

[27] M. E. J. Newman and E. A. Leicht. Mixture models and exploratory data analysis in networks. *PNAS*, 104(23):9564–9569, 2007. doi: 10.1073/pnas.0610537104. URL http://www.pnas.org/cgi/content/abstract/104/23/9564.

[28] Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003. URL http://arxiv.org/abs/cond-mat/0303516.

[29] Manfred Paier and Thomas Scherngell. Determinants of collaboration in European R&D networks: Empirical evidence from a binary choice model perspective. SSRN Working Paper Series No. 1120081, Rochester, NY, July 2008. URL http://ssrn.com/abstract=1120081.

[30] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, June 2005. doi: 10.1038/nature03607. URL http://arxiv.org/abs/physics/0506133.

[31] Josep M. Pujol, Javier Bejar, and Jordi Delgado. Clustering algorithm for determining community structure in large networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74(1):016107, 2006. URL http://link.aps.org/abstract/PRE/v74/e016107.

[32] Joerg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74(1):016110, 2006. URL http://link.aps.org/abstract/PRE/v74/e016110.

[33] Thomas Roediger-Schluga and Michael J. Barber. The structure of R&D collaboration networks in the European Framework Programmes. Working Paper 2006-036, UNU-MERIT, 2006. URL http://www.merit.unu.edu/publications/wppdf/2006/wp2006-036.pdf.

[34] Thomas Roediger-Schluga and Michael J. Barber. R&D collaboration networks in the European Framework Programmes: Data processing, network construction and selected results. *IJFIP*, 4(3/4):321–347, 2008. Special Issue on "Innovation Networks".

[35] Andrew J. Seary and William D. Richards. Spectral methods for analyzing and visualizing networks: an introduction. In Ronald Breiger, Kathleen Carley, and Philippa Pattison, editors, *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, pages 209–228, Washington, D.C., 2003. The National Academies Press. URL http://www.sfu.ca/~richards/Pages/NAS.AJS-WDR.pdf.

[36] S. Wasserman and K. Faust. *Social Network Analysis—Methods and Applications*. Cambridge University Press, 1994.

[37] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–2, June 1998.